Vol.2011-HPC-129 No.23
2011/3/16

複数 GPU 環境における消費電力と性能の関係に関する考察

酒井貴多[†]　　山口実靖[††]

近年，GPU の処理性能が向上し，GPU を汎用計算に用いる GPGPU への注目が高まっている．GPU は処理性能が高いだけにとどまらず，消費電力当たりの処理性能が高いという特徴を持つ．
本稿では，複数 GPU 環境におけるメモリアクセスと演算処理の性能と消費電力の関係について述べる．2GPU を搭載した計算機において使用 GPU 数，スレッド数，ブロック数を変化させてメモリアクセスと演算処理の性能と消費電力の関係を調査したところ，使用 GPU 数の増加に伴い性能と消費電力がともに大きく上昇すること，性能の向上が消費電力の向上を上回るため複数の GPU を使用した方が単位電力あたりの性能が優れていること，演算処理において使用ブロック数が十分に大きくない場合は消費電力が大きいにもかかわらず性能が高くないこと，ブロック数を十分に増加させることにより小さな消費電力の増加により大きな性能向上を実現できることが確認された．

A Study on Performance and Power Consumption in Multiple-GPU Environment

Tota Sakai[†]　and Saneyasu Yamaguchi [††]

GPU has much higher performance than CPU. GPGPU has become an important method to achieve high performance. In addition, GPU has better power consumption/performance ratio. In this paper, we discuss the relation between GPU's performance and its power consumption in multiple-GPU environment. Our experiments demonstrated that multiple-GPU machines have even better power consumption/performance ratio than single-GPU machines.

1. はじめに

描画演算処理装置である GPU (Graphic Processing Unit)は CPU よりも高いピーク性能を持っており，ハードウェアの特性を考慮して処理を行えば非常に高い性能を得ることができる．そのため，GPU を描画演算処理以外の汎用計算に使用する GPGPU (General-purpose computing on graphics processing units)技術に関する研究が盛んに行われている．これらの研究は大きな成果を上げており，例えば 2010 年 11 月におけるスーパーコンピュータランキング TOP500[1)]にて上位 4 システムのうち，日本の TSUBAME 2.0 などの 3 種のシステムが NVIDIA GPU を使用するシステムとなっている．GPGPU がこのような大きな注目を集めている理由の 1 個に，その消費電力があげられる．GPU の消費電力は CPU と比較しても高いが，それを上回る程度で性能が高く，結果的に GPU は電力効率において優れている．省エネルギーの重要性はますます高まっており，今後も GPGPU の重要性は高まっていき，GPU の消費電力に関する考察や低消費電力化も重要になってくると考えられる．

本稿では複数 GPU 環境の消費電力に着目し，メモリアクセスと演算処理における消費電力と処理性能について調査し，考察する．

2. GPGPU

GPGPU は GPU を用いて描画演算処理以外の汎用計算を行う技術である． GPGPU プログラミング環境として，NVIDIA 社が提供している CUDA がある．CUDA を用いることにより，描画演算処理に対する専門的な知識が無くても GPU を用いた汎用計算が可能となる．以下，CUDA GPU に限定して解説を行う．

CUDA GPU は，図 1 の様な構成をしている．すなわち，GPU 内に複数の SM (Streaming Multiprocessor)が存在し，各 SM 内に複数の SP (Streaming Processor)が存在している．これらの SP が処理装置の単位となる．メモリとしては主に VRAM と Shared Memory が存在し，VRAM は全ての SM からアクセスが可能であり，Shared Memory は SM 内に存在し各 SM 内からのみアクセス可能である．VRAM は Shared Memory より容量が多い(数百 MB～数 GB 程度)が，Shared Memory と比較しアクセスに要する時間が長い．Shared Memory は SP からの高速アクセスが可能であるが，容量が少なく(16KB)，キャッシュとして使用されることが多い．つまり，複数回アクセスされるデータを VRAM から Shared Memory にコピーし，Shared Memory に対してアクセスを繰

[†]工学院大学大学院　工学研究科　電気・電子工学専攻
Electrical Engineering and Electronics, Kogakuin University Graduate School
[††]工学院大学　工学部　情報通信工学科
Department of information and Communications Engineering, Kogakuin University

ⓒ 2011 Information Processing Society of Japan

Vol.2011-HPC-129 No.23
2011/3/16

り返すことによりメモリアクセス時間の短縮を実現する．

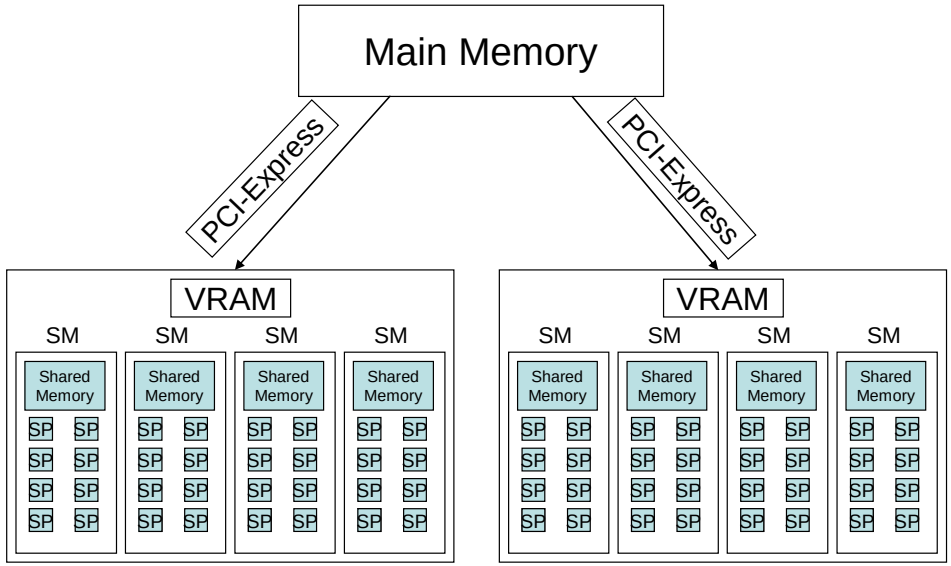

図 1　CUDA GPU の構造

3. 測定環境

本章にて，本研究で行なった GPU 消費電力測定の測定環境について説明する．本実験では，計算機全体の消費電力を測定し，GPU の消費電力について考察する．

測定は，NVIDIA GeForce GTX 460 を使用し，図 2 の様な実験環境を構築し供給電力を測定した．GeForce GTX 460 と CUDA の仕様は表 1　の通りである．

表 1　使用 GPU の仕様

"GeForce GTX 460"	
CUDA Driver Version	3.2
CUDA Runtime Version	3.1
CUDA Capability Major revision number	2
CUDA Capability Minor revision number	1
Total amount of global memory	1072889856 bytes
Number of multiprocessors	7
Number of cores	224
Total amount of constant memory	65536 bytes
Total amount of shared memory per block	49152 bytes
Total number of registers available per block	32768
Warp size	32
Maximum number of threads per block	1024
Maximum sizes of each dimension of a block	1024 x 1024 x 64
Maximum sizes of each dimension of a grid	65535 x 65535 x 1
Maximum memory pitch	2147483647 bytes
Texture alignment	512 bytes
Clock rate	1.35 GHz

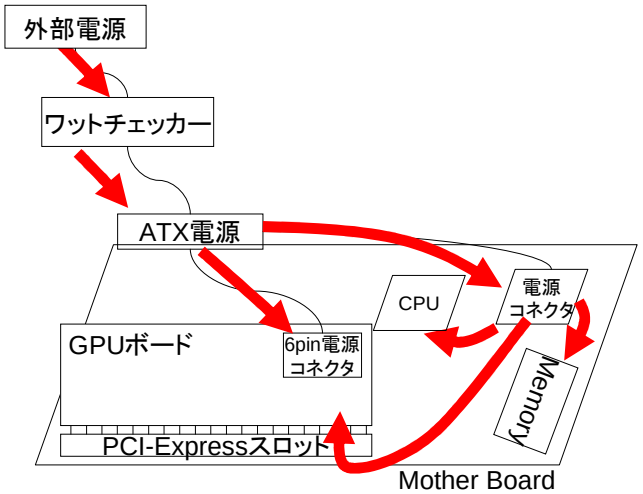

図 2　測定環境

ⓒ 2011 Information Processing Society of Japan

4. 実験

4.1 メモリアクセスの性能と消費電力

GPU には GPU 専用のデバイスメモリである VRAM が搭載されている．本節では，ホストメモリと VRAM 間におけるデータ転送処理および VRAM 間におけるデータ転送処理の性能と消費電力について考察する．

最初に，単一 GPU 環境におけるホストメモリから VRAM へのデータ転送処理，VRAM からホストメモリへのデータ転送処理，VRAM から VRAM へのデータ転送処理の性能と消費電力について述べる．本測定では，CUDA Runtime API の cudaMemcpy を用いてデータ転送を行った．転送量は，400[MB]とした．測定結果を図 3 に示す．これらの性能を単位消費電力当たりの性能に換算したものを図 4 に示す．図内の消費電力は全て PC 全体の消費電力である．これらの結果より，ホストメモリと VRAM 間のデータ転送における転送速度および消費電力は，転送方向に関係無く，ほぼ等しいことがわかった．また，VRAM 間でデータ転送を行なう場合，ホストメモリを介す場合と比較して非常に高い性能が得られると同時に消費電力も高くなることが確認された．単位消費電力当たりの性能で比較すると，VRAM とホストメモリ間の転送より VRAM 間の転送の方が高くなっている．

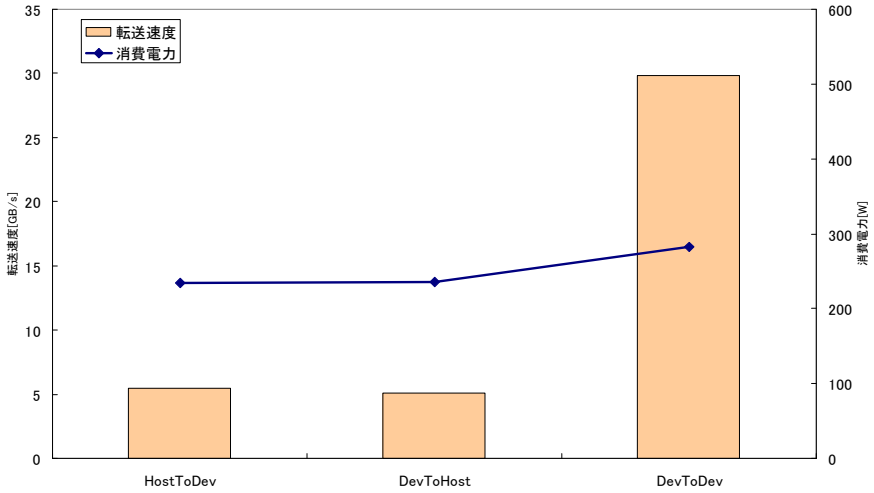

図 3　単一 GPU におけるメモリコピーの転送速度と消費電力

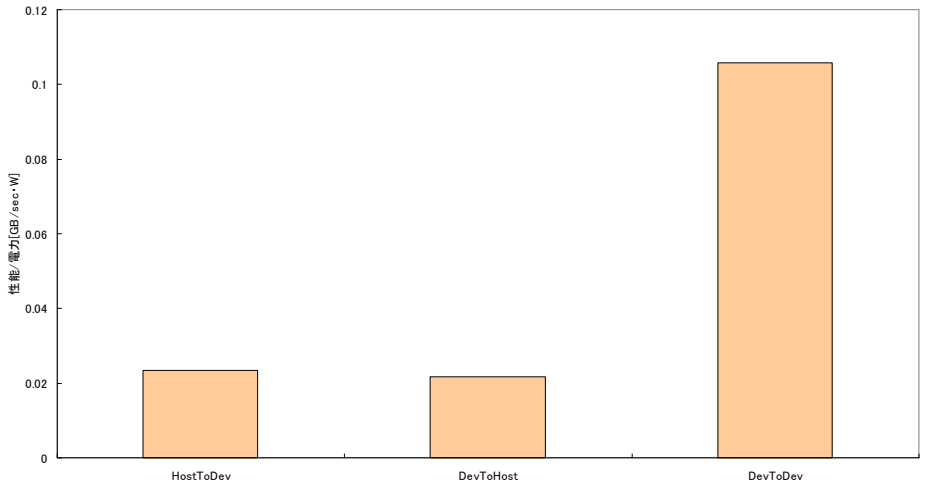

図 4　単一 GPU におけるメモリコピーの単位消費電力当たりの転送速度

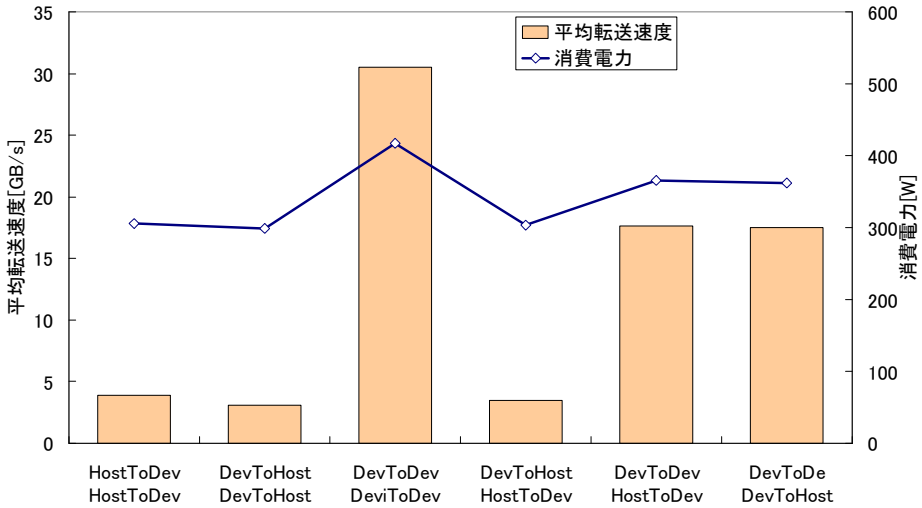

図 5　複数 GPU におけるメモリコピーの転送速度と消費電力

ⓒ 2011 Information Processing Society of Japan

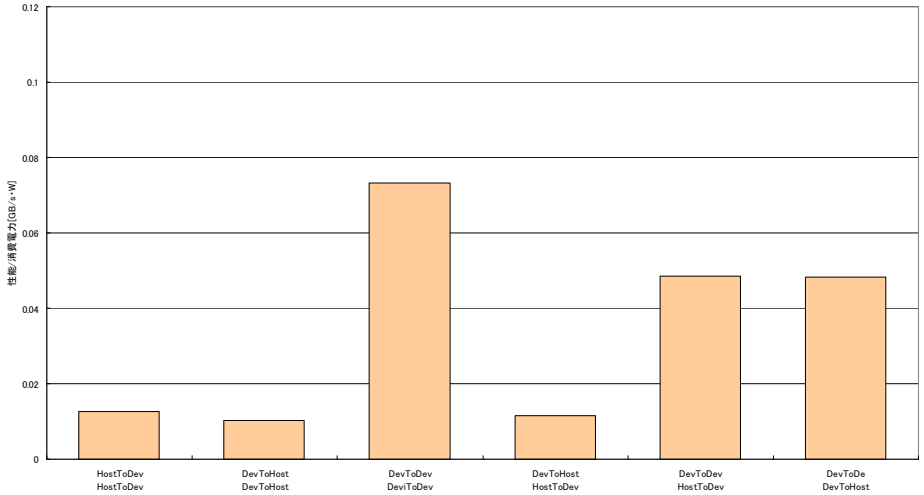

図 6　複数 GPU におけるメモリコピーの単位消費電力当たりの転送速度

同様に，複数 GPU 環境におけるホストメモリ，VRAM のデータ転送処理の性能と消費電力について調査した．本測定でも，CUDA Runtime API の cudaMemcpy を用いてデータ転送を行った．転送量は，400[MB]である．測定結果を図 5 に示す．これらの性能を単位消費電力当たりの性能に換算したものを図 6 に示す．最初に，2GPU で同一のデータ転送処理を行ったときの性能と消費電力の関係について考察する．図より 2GPU で HostToDevice を同時に行った場合と 2GPU で DeviceToHost を同時に行なった場合は，2GPU の平均転送速度が単一 GPU の転送速度よりも低いことが確認できた．また，HostToDevice よりも DeviceToHost の方が平均転送速度の減少が大きいことが分かった．これより，ホストメモリへの書き込みが競合したときの方が，性能劣化が大きくなると考えられる．次に，2GPU で DeviceToDevice を同時に行なった場合について考察する，2GPU の平均転送速度は，単一 GPU の転送速度とほぼ等しいことが確認できた．次に，2GPU で異なるデータ転送処理を行なった場合について考察する．図 5 より，2GPU で DeviceToHost と HostToDevice を同時に行なった場合，2GPU の平均転送速度は，単一 GPU の転送速度よりも減少することが確認できた．2GPU で DeviceToDevice と HostToDevice を同時に行った場合，DeviceToDevice と DeviceToHost を同時に行なった場合の性能は，単一 GPU 実験の性能とほぼ同じになることが確認できた．すなわち，図 5 における DeviceToDevice と HostToDevice の実験におけるそれぞれの転送速度は 29[GB/s]と 5.5[GB/s]であり，単一 GPU 環境とほぼ同一であった．DeviceToDevice と DeviceToHost の実験も，それぞれの転送速度は 29[GB/s]と 5.1[GB/s]であり単一 GPU 環境とほぼ同一であった．

これは，1 つの GPU で DeviceToDevice を行なった場合，もう片方の GPU はホストメモリを専有して使用することができるため，ホストメモリの競合が発生せず転送速度が低下しなかったものと考えられる．

付録に，HostToHost に該当するホストメモリ間でのメモリコピー(CUDA runtime API でない memcpy)と，DeviceToHost を同時に実行したときの性能と，ホストメモリ間でのメモリコピーと HostToDevice を同時に実行したときの性能を示す．

4.2 演算処理の性能と消費電力

使用 GPU 数，スレッド数，ブロック数を変化させて，プロセッサ演算時の性能と消費電力の関係を調査した．行なった処理は Monte Carlo Simulation ある．1 辺の長さが 1 の正方形の中からランダムに座標を選択し，その点が扇形の内部に入る確率を求め円周率を求めた．乱数は発生済みの状態で行い，乱数発生処理は Monte Carlo シミュレーションに含めていない．ブロック数は 1，2，4，7，16，64，256，512，1024，2048，4096，8192 と変化させ，ブロック当たりのスレッド数は 1～1024 まで変化させた．単一 GPU におけるブロック数，スレッド数，性能，消費電力の関係を図 7～図 10 に示す．同一の処理を 2GPU で行なった場合のブロック数，スレッド数，性能，消費電力の関係を図 11～図 14 に示す．

最初に，単一 GPU におけるスレッド数，ブロック数を変化させた場合の性能と消費電力の関係について考察する．図より，単一 GPU ではブロック数が GPU の SM 数である 7 より少ない場合(図 7)，スレッド数を増加させると消費電力が大幅に増加するが，性能の向上が(ブロック数が多い場合と比較して)少ないことが確認できた．そして，7 以上のブロック数を用いた場合(図 8～図 10)，ブロック数の増加に伴い消費電力が緩やかに増加し，性能が大幅に向上することを確認できた．つまり，SM 数以上のブロック数を用いた方が，単位消費電力当たりの性能において優れていることが確認できた．

次に，2GPU の結果について考察する．1GPU の結果と同様に，ブロック数が SM 数よりも少ない場合，スレッド数を増加させると消費電力は大きく増加するが性能の向上は少ないことが分かる．そして，ブロック数が SM 数よりも多い場合，ブロック数の増加に伴い消費電力が緩やかに増加し性能は大幅に向上することが確認された．

次に，1GPU 使用時と 2GPU 使用時の比較を行う．図 7～図 10 と，図 11～図 14 を比較すると，使用 GPU 数を 1 個から 2 個に増加させると，性能は 2.0 倍に向上し，計算機全体の消費電力の増加は 1.7 倍程度にとどまることが確認された．よって，GPU2 個における処理の方が，GPU1 個における処理より単位消費電力当たりの性能が高いことが分かる．

ⓒ 2011 Information Processing Society of Japan

Vol.2011-HPC-129 No.23
2011/3/16

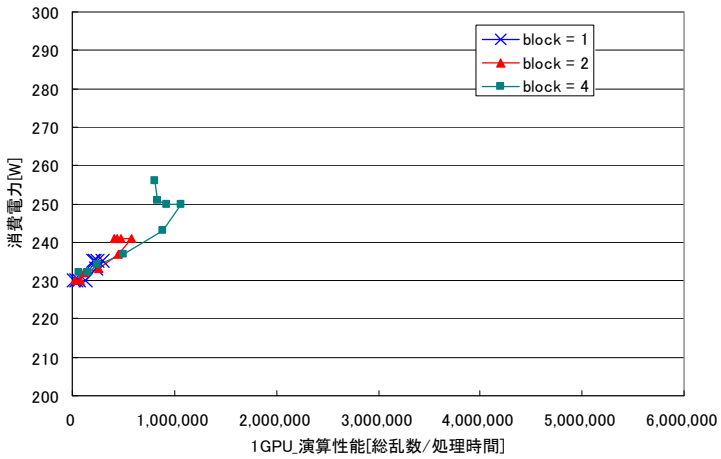

図 7　1GPU における性能と消費電力の関係

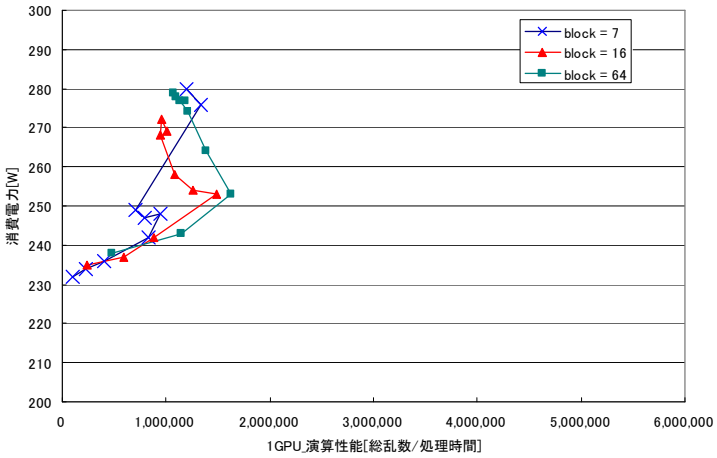

図 8　1GPU における性能と消費電力の関係

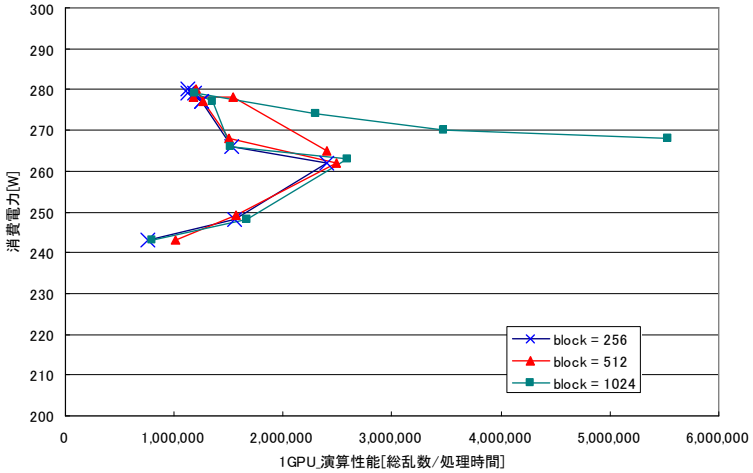

図 9　1GPU における性能と消費電力の関係

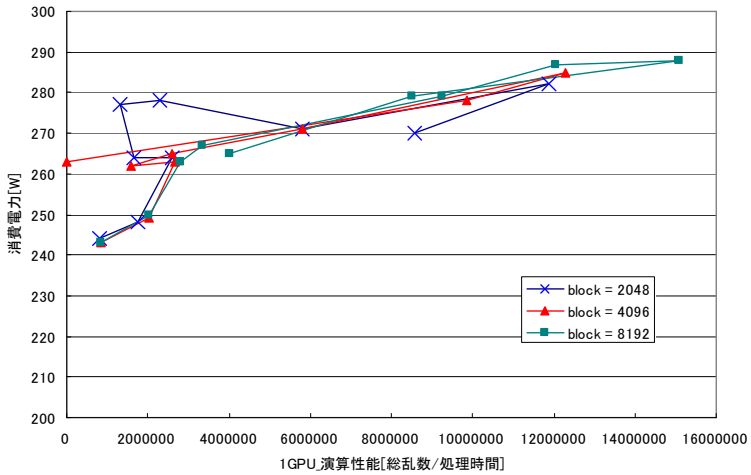

図 10　1GPU における性能と消費電力の関係

ⓒ 2011 Information Processing Society of Japan

Vol.2011-HPC-129 No.23
2011/3/16

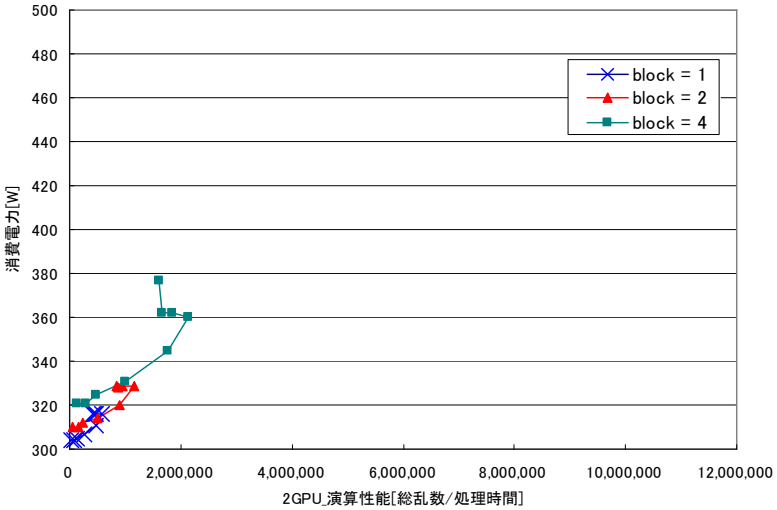

図 11　2GPU における性能と消費電力の関係

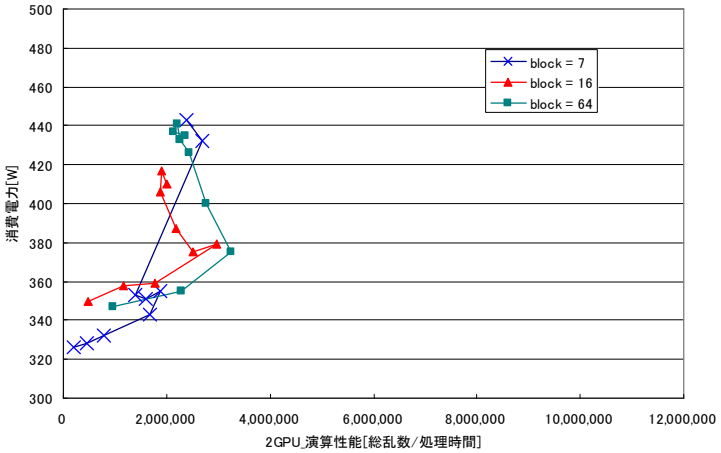

図 12　2GPU における性能と消費電力の関係

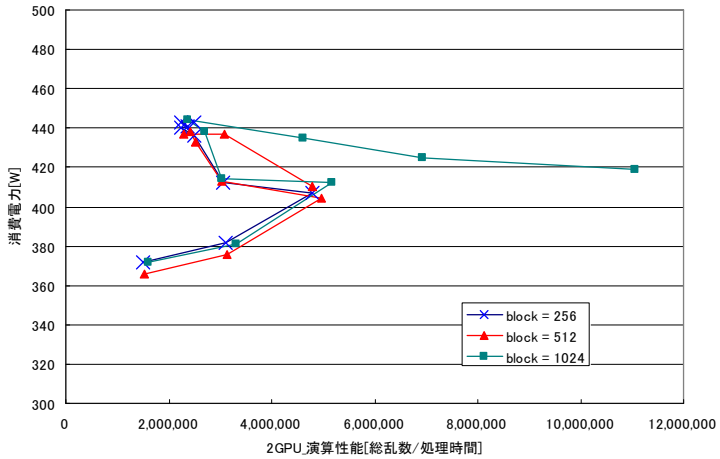

図 13　2GPU における性能と消費電力の関係

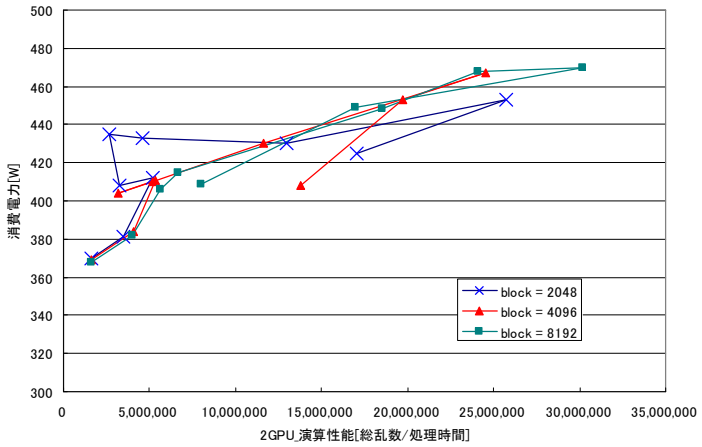

図 14　2GPU における性能と消費電力の関係

ⓒ 2011 Information Processing Society of Japan

5. 関連研究

GPGPU における消費電力に関する研究として，長坂らによる GPU における性能と消費電力の相関性の解析の研究がある [2]．当該研究では，様々な演算プログラムを実行させ，実行プログラム毎に GPU における消費電力が大きく異なることに注目し，GPU プログラムのプロファイラの出力結果と消費電力の関係を調査し，それに基づく消費電力の予測方法を示している．

CPU に代わり GPU を用いることによる省電力に着目した研究として，中村らによるブール自動並列化の研究 [3]や，宇田川らによる GPU を用いた分子動力学法の高速化と省電力化の研究 [4]がある．当該研究では，GPU を用いて高速に分子動力学法の計算を行う方法を示し，GPU を用いることにより単位計算量あたりの消費電力を大幅に削減できることを示している．また，GPU を用いるスーパーコンピュータと，用いないスーパーコンピュータの比較を行い，GPGPU には性能面のみならず消費電力の面にも大きな優位性があることを述べている．

GPU による高速計算手法の確立としては、鈴木らによる高精度演算の並列化 [5]や，宗川らによる配列アライメント処理の高速化の研究 [6]がある．

6. おわりに

本稿では複数 GPU 環境の消費電力に着目し，データ転送処理および演算処理の性能と消費電力の関係について調査を行なった．調査の結果，データ転送処理では VRAM 間におけるデータ転送処理が性能と消費電力の両方が高いことが確認された．また，2GPU でホストメモリを使用する処理を行なった場合，ホストメモリの使用が競合し，性能が低下することが確認された．2GPU でホストメモリを使用しない処理を行なった場合，性能は低下しないことが確認された．また，演算処理時の性能と消費電力の関係について調査を行なった．調査の結果，使用ブロック数が SM 数よりも少ない場合，消費電力が大きいにもかかわらず性能が高くないこと，ブロック数を SM 数よりも十分に増加させることにより小さな消費電力の増加により大きな性能向上を実現できることが確認された．

今後は，省費電力化手法の考察を行う予定である．

謝辞　本研究は科研費（22700039）の助成を受けたものである．

参考文献

1) htp://www.top500.org
2) 長坂 仁，丸山 直也，額田 彰，遠藤 敏夫，松岡 聡: GPU における性能と消費電力の相関性の解析，情報処理学会研究報告 2009-HPC-121(SWoPP 2009) (2009).
3) 中村晃一，林崎弘成，稲葉真理，平木敬: SIMD 型計算機向けループ自動並列化手法，情報処理学会研究報告，Vol.2010-HPC-126,No.10,pp.1-8(2010)
4) 宇田川拓郎，関嶋政和: GPU を用いた分子動力学法の高速化と省電力化，情報処理学会研究報告，Vol.2010-HPC-127,No.5,pp.1-5(2010)
5) 鈴木智博，高精度総和計算と高精度内積計算の GPU のための並列アルゴリズム: 情報処理学会論文誌，Vol.3,No.2,pp.48-56(2010)
6) 宗川裕馬，伊野 文彦，萩原 兼一: 統合開発環境 CUDA を用いた GPU での配列アライメントの高速化手法，情報処理学会研究報告, Vol.2008,No.19(2008-HPC-114),pp.13-18(2008)

付録

HostToHost と DeviceToHost または HostToDevice の性能

HostToHost (memcpy) と HostToDevice を同時に実行したときの性能と，HostToHost と DeviceToHost を同時に実行したときの性能を図 15 に示す．ホストメモリへの書き込みが競合した場合の方が転送速度の減少が大きいことが分かる．

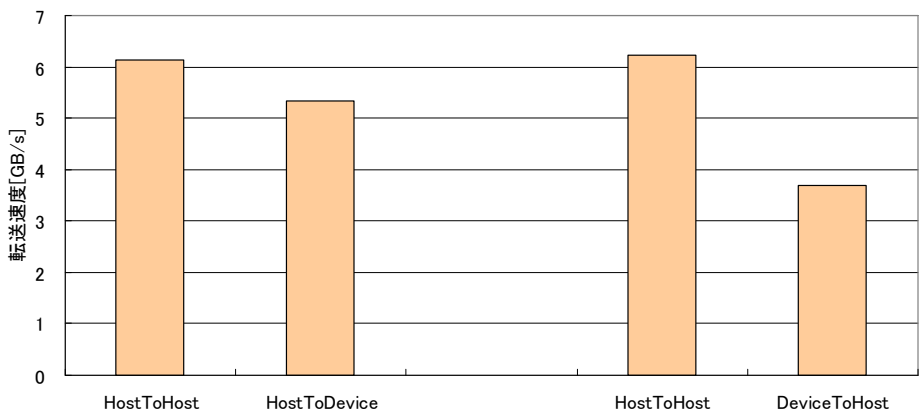

図 15　メモリコピーの転送速度

ⓒ 2011 Information Processing Society of Japan